

Supplementary Methods

Statistical models for sQTL analysis

Here we describe the different models used for sQTL analysis using RNA-Seq data.

1. Linear model (lm)

All previous sQTL analyses took this approach using either the absolute exon read counts [1] or the percentage of the exon reads among all reads in the gene [2]. In our analysis, we used the point estimate of exon inclusion level ($\psi = y/n$) from splice junction reads (Supplementary Figure S1). For individual i , g_{ij} denotes the genotype at SNP_j , then the model is:

$$y_i / n_i = \beta_0 + \beta_j g_{ij} + \varepsilon_i$$

2. Generalized linear model (glm)

Using exon reads directly would be confounded by variation in gene expression levels, making it difficult to distinguish real exon-level splicing QTLs from expression QTLs. Using point estimate of inclusion level (ψ) eliminates the information about the confidence level of the estimate reflected by the read counts that support it.

A natural solution is to use the reads information from both the exon inclusion and skipping isoforms instead of only a point estimate of the exon inclusion level in sQTL analysis. Given the observed splice junction read counts as in Supplementary Figure S1, we can assume that the splice junction reads supporting 2 different isoforms follow a binomial distribution. This leads to a hierarchical model:

$$y_i | \psi_i \sim \text{Binomial}(n_i, \psi_i)$$

$$\psi_i = \text{logit}^{-1}(\beta_0 + \beta_j g_{ij})$$

Where $\text{logit}^{-1}(\beta_0)$ is the baseline exon inclusion level, and β_j is the additive effect of SNP_j . In generalized linear model, this would be to test the null hypothesis $H_0: \beta_j = 0$. We used glm function with logit link function in R package for this model.

3. Dispersion estimate

The simple logistic regression in glm assumes that ψ is correctly modeled and thus: $E(y_i) = n_i \psi_i$, and $Var(y_i) = n_i \psi_i (1 - \psi_i)$. In real data, we would always have variation of the exon inclusion level across individuals in the same genotype group. Thus the observed variance would be larger than the expected variance which leads to overdispersion in the model. We can estimate the dispersion parameter ϕ , where $Var(y_i) = \phi n_i \psi_i (1 - \psi_i)$. The estimate of ϕ can be obtained from the Pearson statistic [3]:

$$\hat{\phi} = \frac{\sum_i (y_i - n_i \hat{\psi}_i)^2 / (n_i \hat{\psi}_i (1 - \hat{\psi}_i))}{m - p}$$

where m is the total number of individuals and p is the number of estimated parameters in the model. When there is no dispersion, ϕ should be equal to 1. To test the overdispersion in the CEU data set, for the top sQTLs (false positive rate $< 1\%$ based on permutations of individual labels) discovered from the glm model, we calculated ϕ and obtained the p-value from the χ^2 distribution with $(m-p)$ degrees of freedom.

4. Generalized linear mixed model prediction of sQTL (GLiMMPS)

To deal with overdispersion, we modeled the extra variance of ψ as a random effect for each individual i in the model with random effects, $u_{ij} \sim N(0, \sigma_{uj}^2)$ [4]. Let $u_{ij} = \sigma_{uj} z_{ij}$, then $z_{ij} \sim N(0,1)$, β_j denoting the fixed effect for SNP j , which leads to a generalized linear mixed model:

$$\psi_i = \logit^{-1}(\beta_0 + \beta_j g_{ij} + \sigma_{uj} z_{ij}).$$

Thus the joint likelihood for β , σ_{uj} is given by:

$$L(\beta, \sigma_{uj}) = \prod_{i=1}^m \binom{n_i}{y_i} \int \frac{\exp\{(\beta_0 + \beta_j g_{ij} + \sigma_{uj} z_{ij})\}^{y_i}}{[1 + \exp\{(\beta_0 + \beta_j g_{ij} + \sigma_{uj} z_{ij})\}]^{n_i}} N(z_{ij}) dz_{ij},$$

where $N(\cdot)$ is the standard normal density. Laplace approximation was used for the parameter estimations and a likelihood ratio test was used to obtain the P-values for the fixed effect β_j for each SNP j .

Simulations

To compare the power and false positive rate of the various methods we discussed, we simulated the splice junction reads by mimicking the characteristics of the experimental data where possible. For each simulation, we simulated a sample of $m=50$ individuals which is close to the sample size of the experimental data. We simulated 10,000 replicates for each combination of parameters. To examine the effect of sequencing depth, we simulated the total number of splice junction reads (n) with mean value (\bar{n}) equals to 5, 10, 20, 50, and 80, respectively.

Because the total splice junction reads for each exon vary across individuals and they do not follow simple Poisson distribution, we used a Gamma-Poisson hierarchical model. First we generated the mean of Poisson distribution for each \bar{n} value: $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, where the shape parameter $\alpha = \bar{n}\beta$ and the rate parameter $\beta = 0.1$, which is the mean from the model fitting on experimental data. Then we generated n_i for individual i using Poisson distribution $n_i \sim \text{Poisson}(\lambda_i)$. After that, we simulated the genotypes (g_{ij}) for a random SNP j with the minor allele frequency (MAF) modeled by an exponential decay distribution with minimum MAF = 0.05 to obtain 3 genotypes AA, AB, BB. Then we simulated the baseline exon inclusion level ($\bar{\psi}_0$) for genotype AA according to a distribution Beta(0.35, 0.81) that fits the observed pattern in the top sQTLs (false positive rate $< 1\%$ based on permutations of individual labels) identified by GLiMMPS from the

CEU data set (Supplementary Figure S4a). We simulated the difference of mean exon inclusion level of BB from AA ($\Delta\psi$) based on the mean and variance from the observed top sQTLs in the CEU data set (Supplementary Figure S4b). Assuming an additive model, we can obtain the mean ψ for genotype AB as well. Then for each genotype group, a noise was added according to normal distribution with the standard deviation sampled from the distribution observed in the top sQTLs in the CEU data set (Supplementary Figure S4c). After obtaining ψ_i for each individual, the inclusion isoform splice junction read counts were generated as: $y_i \sim \text{Binomial}(n_i, \psi_i)$.

Note that these simulations assume a normal distribution of ψ in each genotype group with the mean, genotype effect and noise mimicking the distributions from top sQTLs in the CEU data set. This is different from the logistic regression link function applied in GLiMMPS, which avoids biasing the simulation results in favor of our model.

Using simulation data with equal mean inclusion levels for three genotype groups and with various read depth, we evaluated the false positive rates (Type I errors) at 5% significance level. We also computed the receiver operating characteristic (ROC) curve by combining all the simulated data with null effects and various SNP effects.

sQTL analysis and estimation of FDR

For both CEU and CEU2 data sets, using each of the statistical models (lm, glm and GLiMMPS) described above, we carried out the analysis for each AS event using SNPs within 200kb of the target exon. To estimate the false discovery rate, we used the same approach as in [2]. For each exon, we permuted the individual label 10 times, recalculated the P-values, and recorded the minimum P-value for each exon for each permutation. This set of minimum P-values serves as the empirical null distribution for the P-values (denoted as p_0). We then compared the true distribution of minimum P-values (denoted as p_1) to this null distribution to obtain the estimate of FDR. For example, for FDR = 0.1, we found a P-value cutoff z such that $P(p_0 < z) / P(p_1 < z) = 0.1$, where $P(p_0 < z)$ is the fraction of minimum P-values from permutation less than z and $P(p_1 < z)$ is the fraction of minimum P-values from observed data less than z . Partitioning the 10 permutations into 5 different sets yielded similar estimates of the FDR curve, suggesting our permutation procedure is sufficient to capture the null distribution genome-wide (Supplementary Figure S5).

References

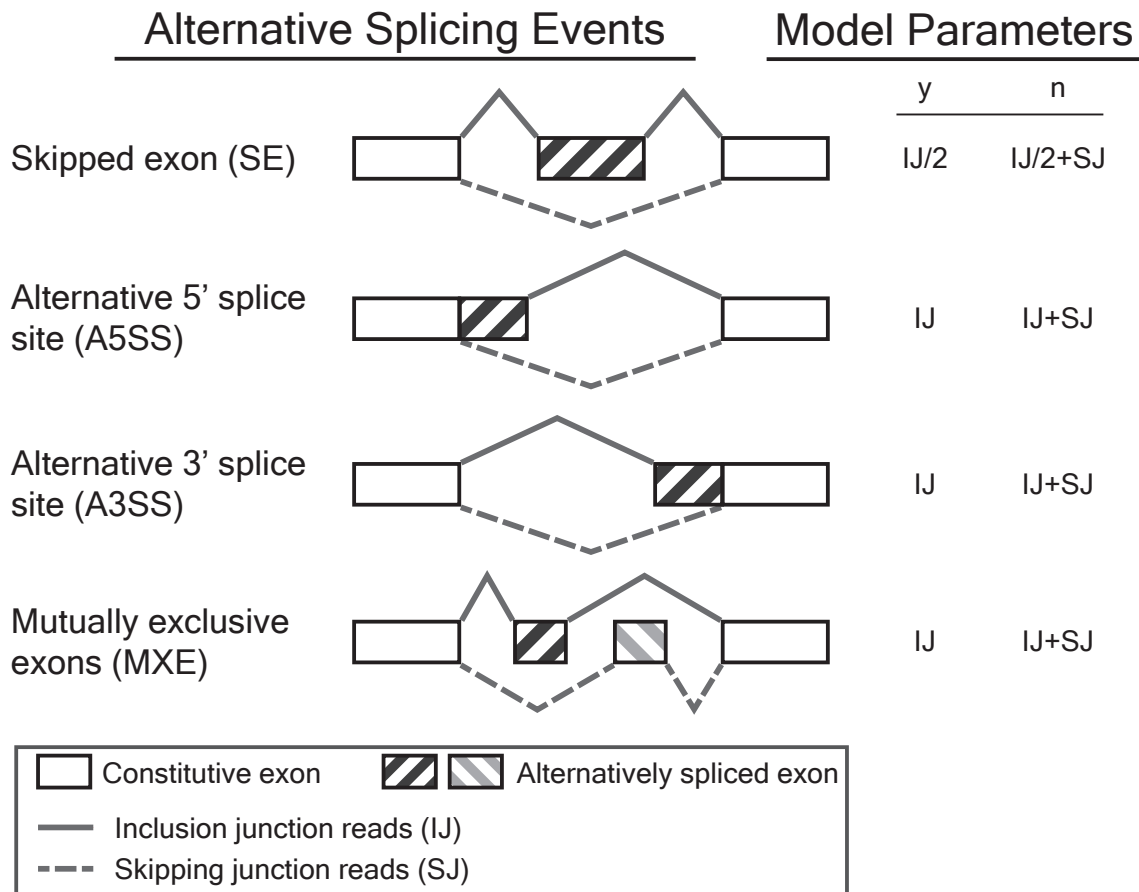
1. Montgomery S, Sammeth M, Gutierrez-Arcelus M, Lach R, Ingle C, Nisbett J, Guigo R, Dermitzakis E: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464**:773-777.
2. Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard J: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768-772.
3. McCullagh P, Nelder J: *Generalized Linear Models*. Chapman and Hall; 1989.

4. Browne WJ, Subramanian SV, Jones K, Goldstein H: **Variance partitioning in multilevel logistic models that exhibit overdispersion.** *Journal of the Royal Statistical Society Series a-Statistics in Society* 2005, **168**:599-613.

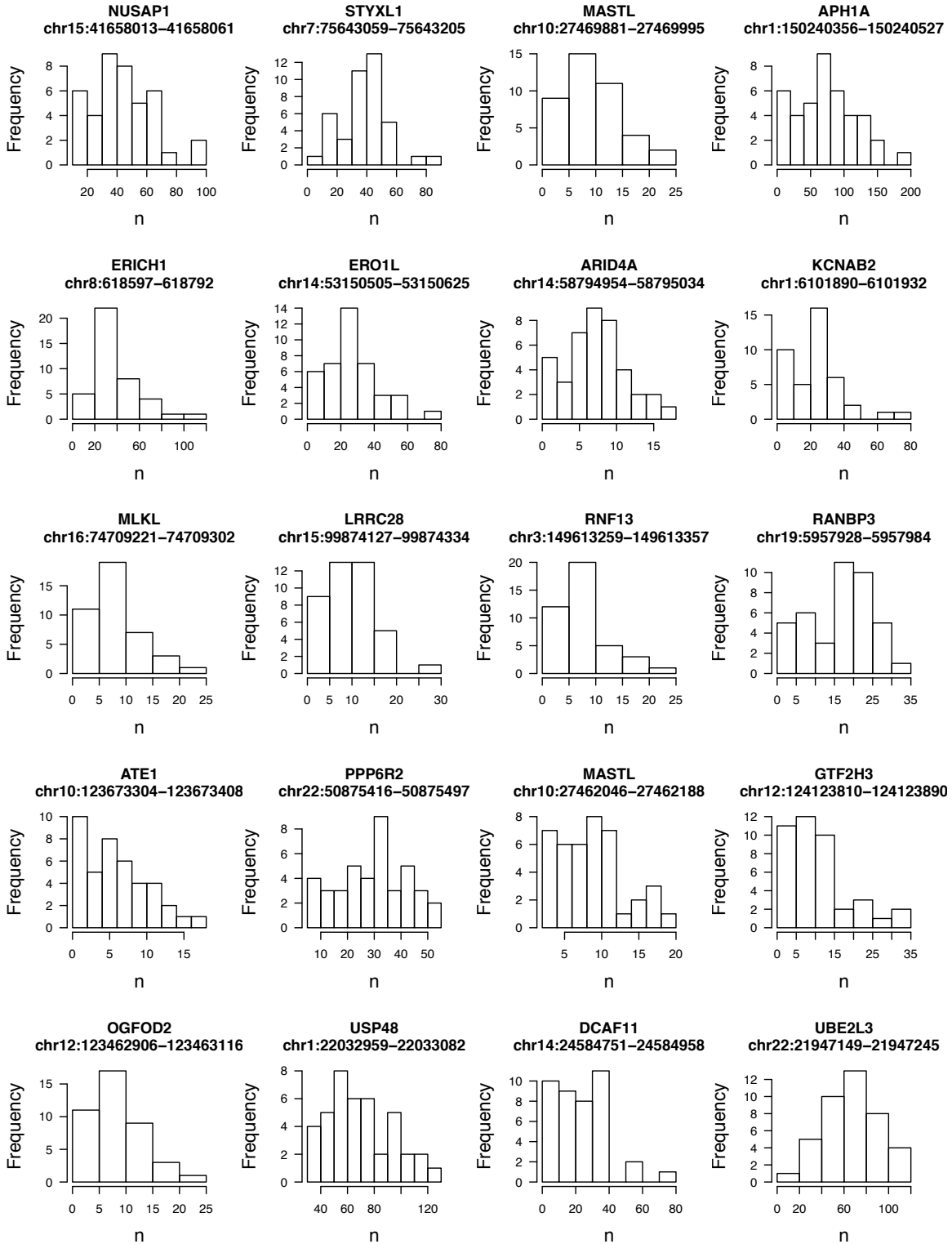
Supplementary Figures

Supplementary Figure S1. Four types of alternative splicing events analyzed.

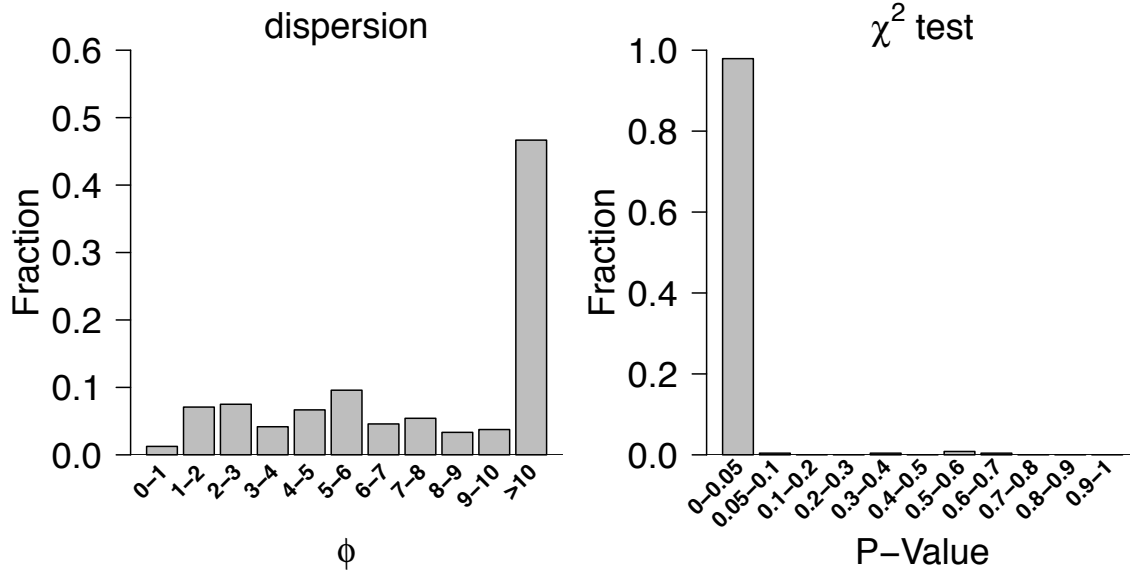
Schematic diagrams of different types of alternative splicing events analyzed. The main target exon, mutually exclusively spliced exon, and neighboring constitutive exons are illustrated in black hatched box, grey hatched box, and open boxes. The exon inclusion and skipping splice junctions of the target alternatively spliced exon are depicted on the top and bottom as solid and dashed lines, respectively. Parameters used in the model are illustrated on the right, with y representing the effective read count from the exon inclusion isoform and n representing the total effective read count from both the exon inclusion and skipping isoforms.



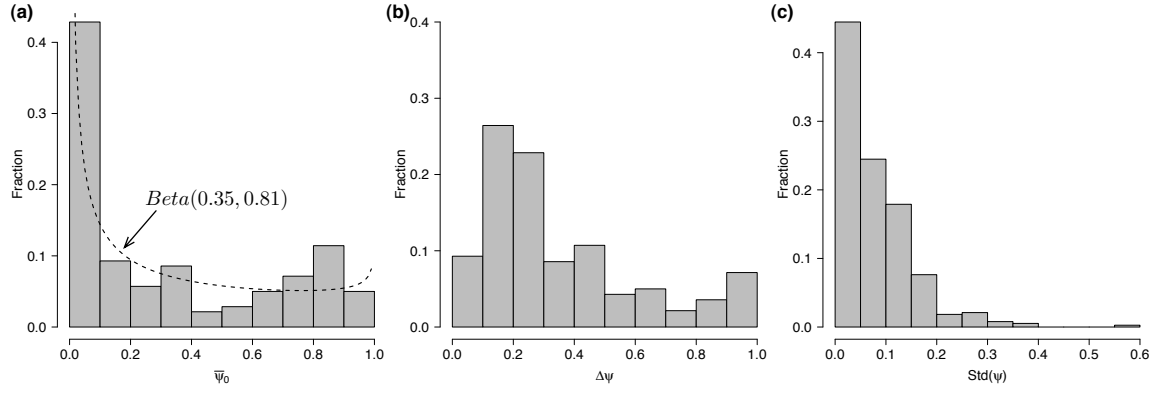
Supplementary Figure S2. Distribution of total splice junction read counts (n) across individuals in the CEU RNA-Seq data set. 20 randomly selected alternatively spliced exons from the CEU data set are shown in the plot.



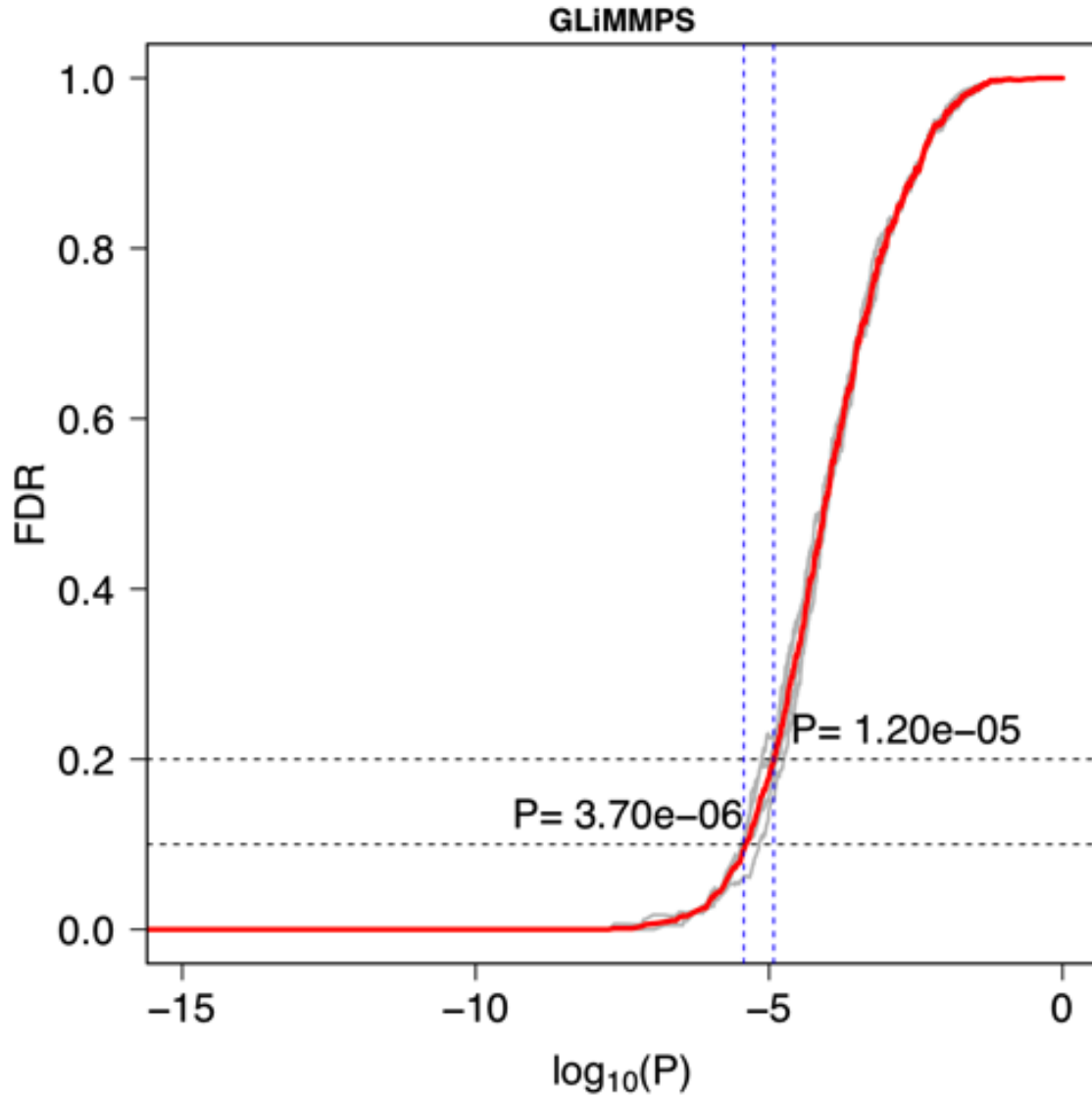
Supplementary Figure S3. Dispersion parameter estimate from top sQTLs (Type I error < 1%) from the glm model on the CEU data set. The estimate of the dispersion parameter ϕ for top sQTLs (false positive rate < 1%) from the glm model was obtained from the Pearson statistic, described in detail in Supplementary Methods. The distribution of the p-values for overdispersion of ϕ is shown on the right.



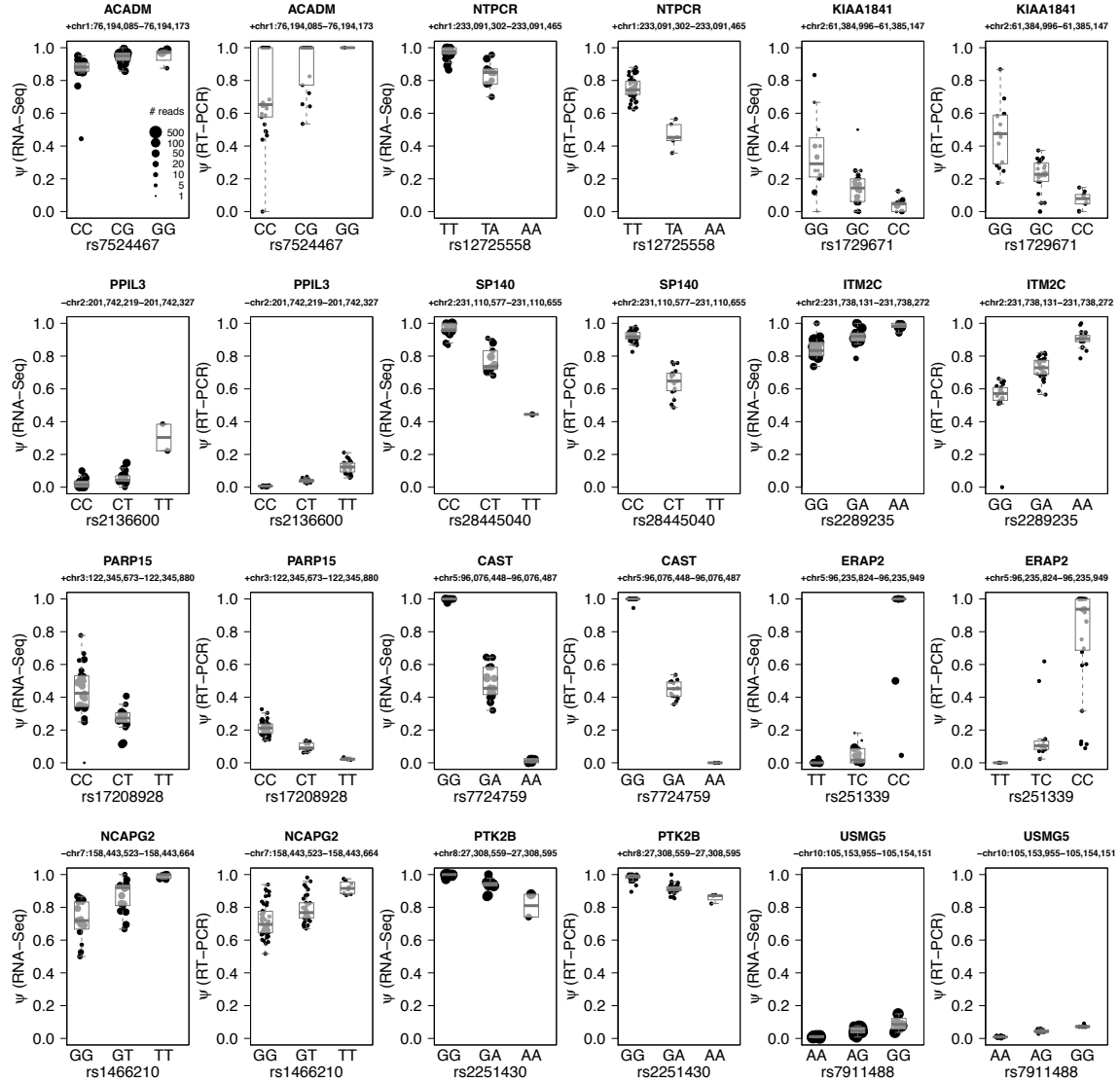
Supplementary Figure S4. Distributions of model parameters for top GLiMMPS sQTLs (false positive rate < 1%) in the CEU data set. (a) Distribution of the mean exon inclusion level ($\bar{\psi}_0$) in the baseline genotype group. The best Beta distribution fitted to the data is shown in the dashed line. **(b)** Distribution of the mean difference of exon inclusion levels ($\Delta\psi$) between the 2 homozygous genotypes. **(c)** Distribution of the standard deviation of exon inclusion level ($\text{Std}(\psi)$) of each genotype group.

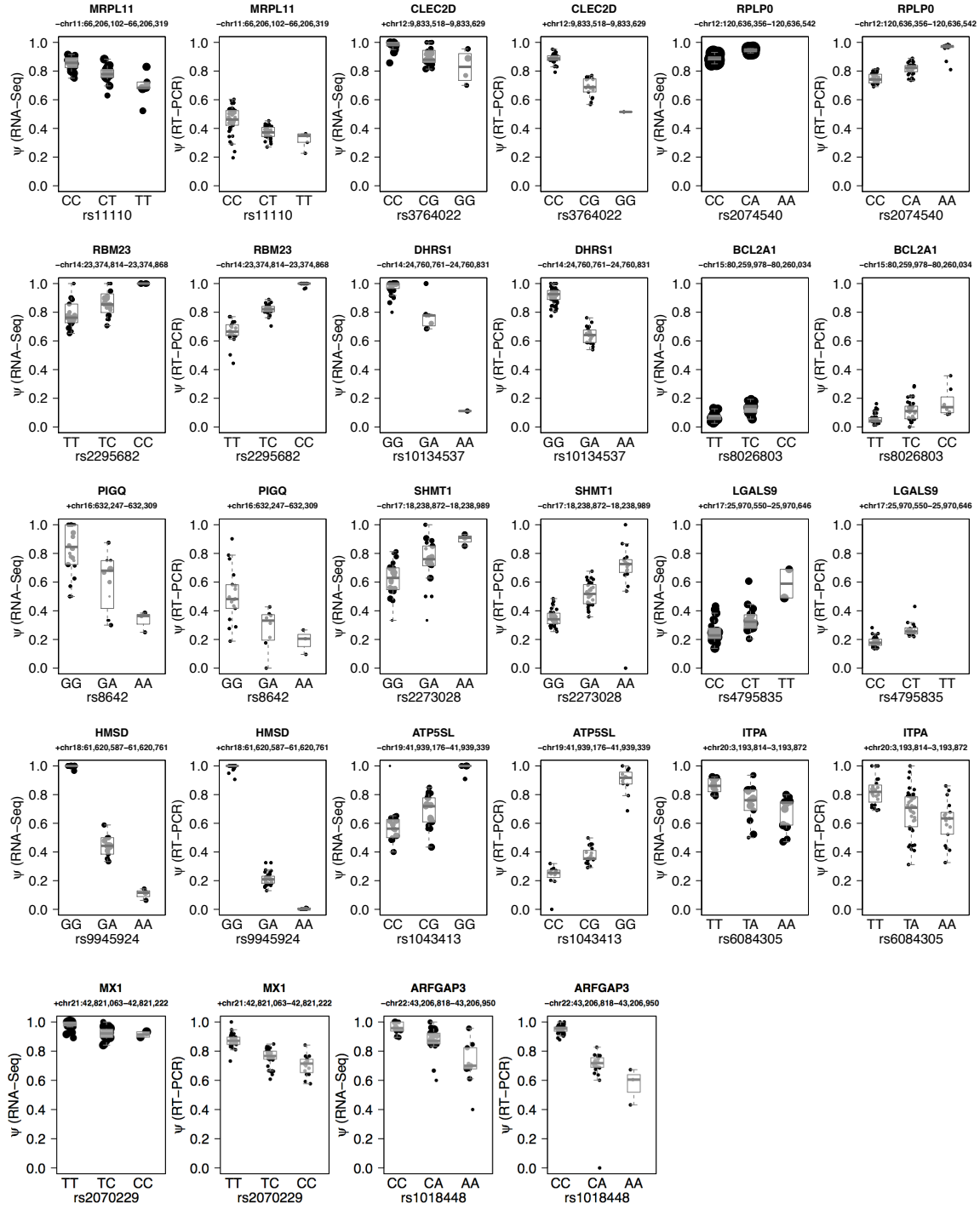


Supplementary Figure S5. FDR estimate based on the most significant P-value for each exon from the permuted data and observed data. The red line represents the FDR estimate from all 10 permutations. Five grey lines represent FDR estimates based on two permutations per line.

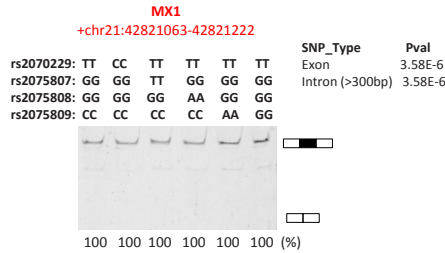
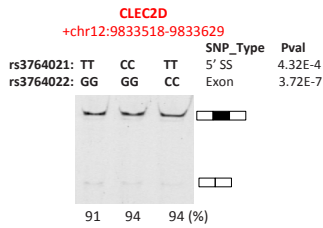
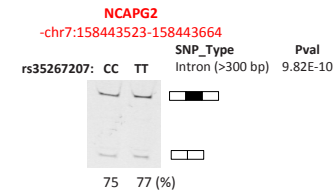
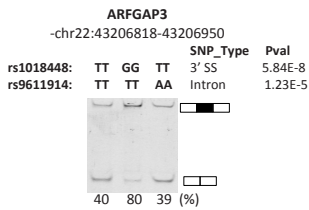
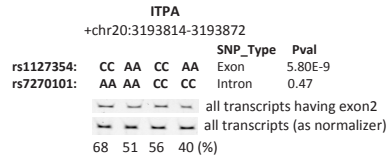
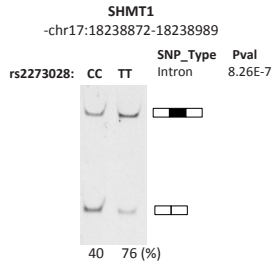
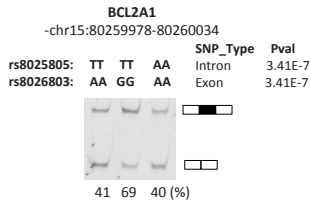
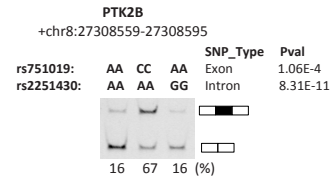
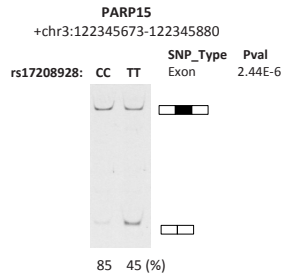
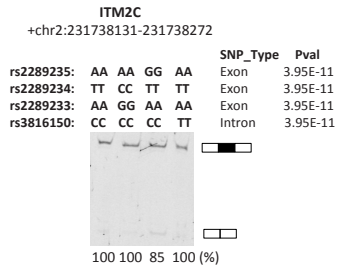
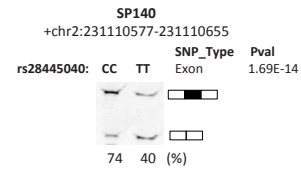
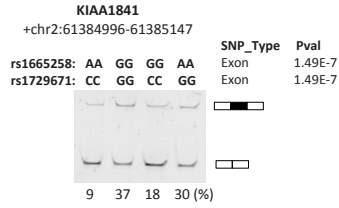
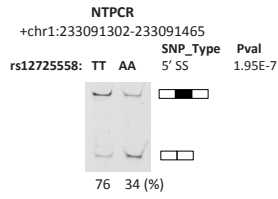


Supplementary Figure S6. Boxplots showing the correlation of exon inclusion levels (ψ) from RNA-Seq and RT-PCR with the most significant HapMap3 SNP. For improved illustration, individuals from the same genotype group are jittered on the X-axis, and semi-transparent box plots showing the median and quartiles are overlaid on top of the dot plot for each exon. In the RNA-Seq plot, the dot size for each individual is proportional to the total number of splice junction reads for the individual.

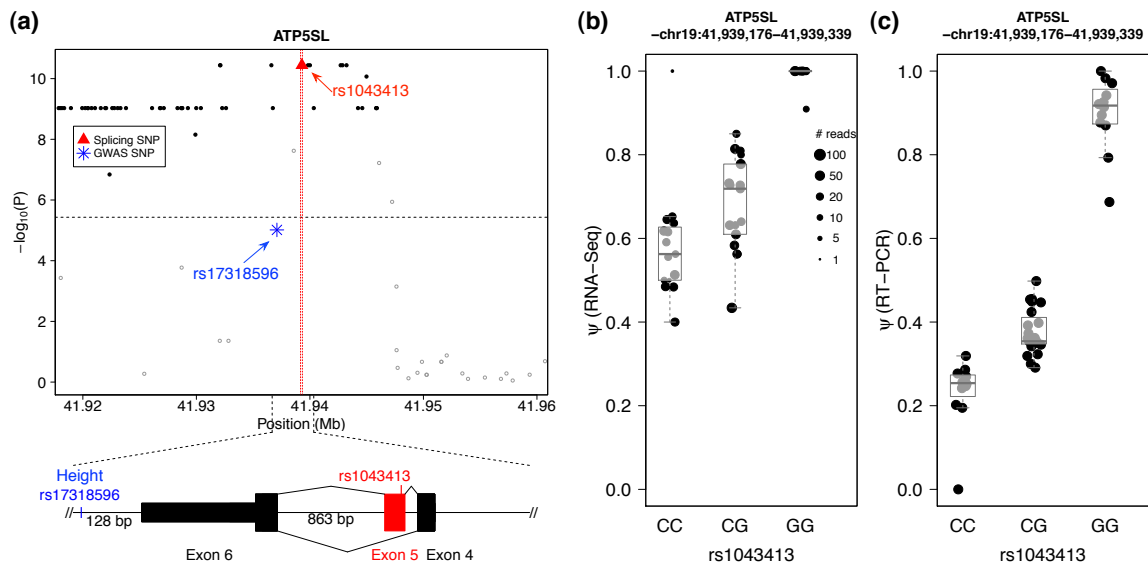




Supplementary Figure S7. Minigene experiments for testing causal sQTL SNPs. For each sQTL target exon, the gene symbol and exon coordinate in hg19 (with start position 0-based and end position 1-based) are shown on top. For each tested SNP, the dbSNP ID, genotypes, SNP annotation type, and GLiMMPS P-value are listed. The estimated exon inclusion levels of different minigene constructs are indicated below each gel picture. Each column represents the minigene construct corresponding to a particular genotype as indicated above the gel picture. The exon inclusion level is calculated from fluorescently labeled RT-PCR data as the intensity of the exon inclusion bands(s) over the total intensity of all exon inclusion and skipping bands. The last 4 exons marked in red have no tested SNPs causing more than 10% change in minigene exon inclusion levels. Universal primer sequences were used to test the inclusion level of target exon with the exception of *ITPA*. The *ITPA* minigene construct contains two adjacent exons (target exon 2 and the downstream exon 3) of *ITPA*. Both exons are alternatively spliced and are similar in size. Based on PCR product size alone, we cannot distinguish the transcript with target exon (exon 2) from the transcript with the downstream exon (exon 3), so we used target exon specific primers for RT-PCR analysis.



Supplementary Figure S8. sQTL signal overlapping with GWAS signal near gene *ATP5SL*. (a) The distribution of GLiMMPS P-values around the sQTL exon (exon 5) in gene *ATP5SL*. The black horizontal dashed line reflects the 10% FDR cutoff and red vertical lines mark the location of the sQTL exon. SNPs in linkage disequilibrium ($r^2 > 0.8$) with the GWAS SNP rs17318596 for height (blue asterisk) are shown in solid black dots, while other SNPs are shown in grey circles. The causal splicing SNP in exon 5 is shown in red triangle. Exon-intron structure is shown in the bottom with the GWAS SNP (rs17318596) and the causal splicing SNP (rs1043413) marked at corresponding locations. (b) Boxplot showing the significant association of rs1043413 with exon inclusion level (ψ) of the *ATP5SL* exon 5 estimated by the CEU RNA-Seq data set. The size of each dot is scaled by the total number of splice junction reads for that individual. (c) The same boxplot using exon inclusion level (ψ) measured by quantitative RT-PCR.



Supplementary Figure S9. sQTL signal overlapping with GWAS signal near gene *ITPA*. (a) The distribution of GLiMMPS P-values around the sQTL exon (exon 2) in gene *ITPA*. The black horizontal dashed line reflects the 10% FDR cutoff and red vertical lines mark the location of the sQTL exon. SNPs in linkage disequilibrium ($r^2 > 0.8$) with the GWAS SNP rs11697186 for response to hepatitis C treatment (blue asterisk) are shown in solid black dots, while other SNPs are shown in grey circles. The causal splicing SNP in exon 2 is shown in red triangle. Exon-intron structure is shown in the bottom with the GWAS SNP (rs11697186) and the causal splicing SNP (rs1127354) marked at corresponding locations. Because exon 2 and 3 are predominantly spliced together, both are colored in red. (b) Boxplot showing the significant association of rs1127354 with exon inclusion level (ψ) of the *ITPA* exon 2 estimated by the CEU RNA-Seq data set. The size of each dot is scaled by the total number of splice junction reads for that individual. (c) The same boxplot using exon inclusion level (ψ) measured by quantitative RT-PCR.

